Irina et al., J Proteomics Bioinform 2017, 10:1
DOI: 10.4172/jpb.1000420

Journal of
Proteomics & Bioinformatics

Review Article OMICS International

Computational Errors and Biases in Short Read Next Generation Sequencing

Irina Abnizova[1*], Rene te Boekhorst[2] and Yuriy L Orlov[3,4]

[1]Wellcome Trust Sanger Institute, Cambridge, UK
[2]University of Hertfordshire, Hatfield, UK
[3]Institute of Cytology and Genetics, Novosibirsk, Russia
[4]Novosibirsk State University, Novosibirsk, Russia

Abstract

Next generation sequencing technologies produce an astronomical amount of useful data, but also artefacts and errors. Some of these errors may mimic true biological signals, such as mutations, and therefore may invalidate conclusions.

In next generation sequencing, two types of errors may occur: experimental and computational. Computational errors are those that stem from the digital post-processing of sequenced samples, and are the main subject of this paper. Post-processing involves procedures such as quality-scoring, aligning, assembling, variant calling, genotyping and error-correction of the data. This paper is about post-processing errors and computational methods to detect and correct them.

Keywords: NGS; Statistical biases; Sequencing errors; Post-processing; Quality control

Introduction

The original Sanger sequencing method [1,2] is referred to as a first-generation DNA sequencing technology. The next generation sequencing technologies (NGS) [3] include: (i) Second generation sequencing, the massive parallel sequencing of relatively short DNA fragments [4]; and (ii) Third generation sequencing, in which single DNA molecules hence much longer fragments are the subject of sequencing [5].

In this paper we will focus on second generation DNA sequencing, and will omit the term 'second generation' while mentioning NGS further.

The Sanger method differs from the NGS in, among other things that it works with relatively large fragments which simplifies assembling. Despite the fact that it is laborious, and therefore time consuming and expensive, the Sanger method is still respected as the most reliable technique and hence functions as the 'gold standard' [6].

This implies that in spite of its sophisticated and elaborated sequencing machinery, the much faster and cheaper NGS technologies are still prone to mistakes that may lead to incorrect conclusions. Artefacts generated during library preparation, in particular as side effects of the Polymerase Chain Reaction (PCR), introduce artificial mutations [7] and sequencing bias. The latter arises because the nucleotide composition of particular regions of the genome may make them less likely to be duplicated depending on the parameter setting of the cloning process. The consequence is that certain parts of the genome are better covered by fragments than others. Ideally, this coverage should be homogeneous, i.e., the counts of nucleotides (from copied fragments) should be uniformly distributed over the positions in the reference genome. Unfortunately, this is often not the case [8]: Particular areas of the genome might be underrepresented because of the sequence complexity and/or function, while other areas might be overrepresented, e.g. repetitive DNA.

Box 1: NGS library construction

Sequencing involves the shearing of DNA into numerous fragments. Originally, restriction enzymes were used to cut off specific parts that were stored in dedicated strains of bacteria or bacteriophages (as "BAC libraries") so that they could be cloned, sorted and fragmented again. Although this is no longer done in current sequencing technology, the name "library" has stuck and is now used in a more general sense for the collection of DNA fragments that has undergone laboratory treatments (including cloning them into a large number of copies to boost the source material by means of the Polymerase Chain Reaction, PCR) to make them suitable for the actual sequencing on instruments specially devised for this purpose.

NGS technologies consist of shearing DNA molecules into collection of numerous small fragments, called a 'library', and their further extensive parallel sequencing. These sequenced overlapping fragments (their fixed length ends actually) are assembled into contiguous strings. The contiguous sequences are in turn further assembled into genomes for further scientific analysis.

DNA sequencing is essential for establishing similarities and spotting deviations (as in mutation screening) between the genomes of individuals and taxa. Hence, its results are used as well to compare genomes from forensic, ecological and evolutionary perspectives as to identify genetic aberrations that might be involved in the aetiology of certain diseases.

We refer to [9-17] for the classification and detailed characteristics of NGS platforms. In spite of sophisticated and elaborated experimental sequencing parts, it is accepted that, strikingly, one of the main challenges in NGS is the digital processing of the big data [3,18].

***Corresponding author:** Irina Abnizova, Wellcome Trust Sanger Institute, Cambridge, UK, Tel: 01223330385; E-mail: ia1@sanger.ac.uk

Received December 26, 2016; **Accepted** January 16, 2017; **Published** January 26, 2017

Citation: Abnizova I, te Boekhorst R, Orlov Y (2017) Computational Errors and Biases in Short Read Next Generation Sequencing. J Proteomics Bioinform 10: 1-17. doi: 10.4172/jpb.1000420

The NGS data processing is arranged in a set of sequential steps, called a pipeline. A typical post-sequencing NGS pipeline [11,12] consists of:

1) Quality control of raw sequence reads;
2) Aligning to a reference genome/assembly;
3) Post-alignment quality control and re-calibration;
4) Identification of mutations (variant calling and genotyping);
5) Post-variant call/genotyping quality control;
6) Data storage and compression.

If no reference exists for the sequenced genome, step 2 may be substituted (or combined) by a de-novo genome assembly step [17]. Each step incorporates some error-correction procedures [10,19].

Box 2: Post-processing pipelines information

There are several good reviews about NGS computational post-processing frameworks [4-12]. There are also good sources online [16,18,20] which can help researchers to create their own pipelines, monitor their data and chose a bioinformatics tools to do so. There are even 'meta' pipelines [13,14] that have been developed, which offer tools to build up customised pipelines.

For each of the first five steps above we will address: i) What they are and aim at; ii) How they work, and iii) The various ways they can be applied, their problems and best practices to solve them. We do this by providing a brief summary of the methodology, presenting an overview of the available tools and a brief assessment of their strengths and weaknesses. We will also review error models and NGS simulation in the section 6 of the paper.

Quality Control (QC) of raw sequence reads

For any platform, initial raw digital outputs of DNA sequencing are nucleotide base calls and their qualities. Base calls and their qualities are usually stored conventionally in the form of FASTQ [21], or BAM/SAM [22] formatted files, and is the input for the majority of a post-processing pipelines.

Box 3: Base calls, their qualities, and reads

The raw digital output of DNA sequencing consists of a series of assessed nucleotide identities ("*base calls*") from a restricted part of the cloned fragments (e.g. the *k* first and last nucleotides, where *k* depends on the NGS technology) called *reads* and the *qualities* of those base calls. A base call quality ("*Q*") corresponds to the probability that a base call deviates from the identity of the corresponding nucleotide in the reference genome [23].

The quality of a base-call may depend on the quality of the signal used in the recognition of a nucleotide, which typically involves the intensity of the identifying fluorescence released by reagents during the sequencing In Illumina, signal quality is measured by a combination of metrics, one of which ("purity") captures the unambiguity of fluorescent intensities. Because sequencing in Illumina is done in cycles (in each cycle one base is called, thus the number of cycles is equal to the read length *k*) and reagents may lose their vigour over time, signal quality, and therefore base-call, quality could be affected by the duration of the run. The decrease of signal quality, and hence of base call quality, with increasing cycle number has been established by a number of studies [24].

The relationship between base-call quality and signal quality is, however, not always straight forward. For instance, although artificial mutations induced during library preparation will show up as discrepancies from the reference genome and by all means are miss-identifications, they are not necessarily base calls of low signal quality.

Compared to Sanger sequencing, NGS technologies are challenged by shorter sequence read length, higher base call error rate, platform/instrument/sample specific artefacts [25], and often low uneven coverage [26,27]. Short reads, in turn, result in limited ability to sequence repetitive DNA [28]. These features lower an accuracy of NGS downstream analysis (e.g. mutation detection and de-novo assembly) by introducing sequencing biases and errors that might lead to incorrect interpretation of data.

A quality control of raw sequence reads is an initial check of the soundness and usefulness of the input. Many of the artefacts brought about by flaws in library preparation and sequencing only become apparent at later stages of the pipeline, but some of them can be detected by QC of raw sequence reads. Therefore it is very important to QC raw sequence reads before further analysis.

The main metrics used in QC of raw sequence reads to characterise artefacts are shortly described below. These metrics reflect library preparation and sequencer's performance, but not post-processing quality.

Total read count should be counted after PCR/optical duplicates removal, which might inflate count's value. It reflects general library usefulness, and should be large enough for a statistical significance of results. The Q value distribution should be skewed towards high quality base calls (majority Q30+), as shown in the Figure 2.

In the Figure 1 one can see an example of a Q distribution for HiSeq v4 Illumina release, Phix spiked-in control, and one lane. There are only five quality bins in this release, so there are 5 vertical bars for Q 10,22,27,33 and 37. The reads here are 100 bp long. One can see that the majority of called bases are of higher quality here, S33+. Generally, to ensure that a data is useful, at least a half of it should be higher than Q30.

A low quality bases arise mostly because of sequencer's biases and imperfections [29,30]. They can add unreliable sequences to the dataset, and lead to false interpretation of data.

Low quality reads or bases at the end of reads are trimmed [31,32], so low quality and possibly wrong called data will not confuse further analysis. However, there are special conditions on an error correction, namely it can be applied only to homogeneous and high-coverage data,

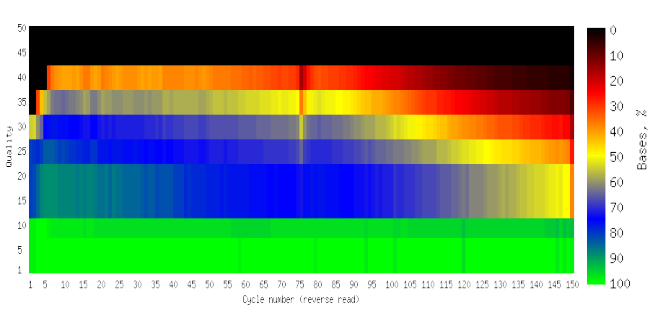

Figure 1: Example of quality values decline per cycle, read length 150 bp, reverse read, Illumina HiSeq. From WTSI quality control web page.

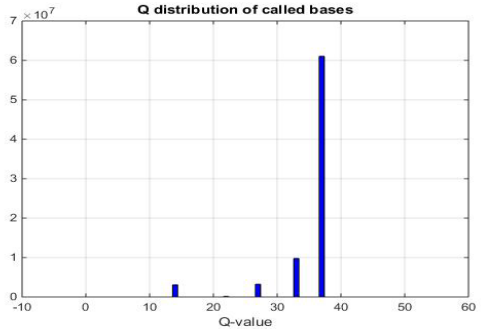

Figure 2: Q-value distribution of all called bases, Illumina HiSeq v4. X-axis is for Q-values. Y-axis is amount of bases called with these values.

which is often not the case for NGS [26,27], especially for the Whole Genome Sequencing (WGS) and GC-extreme regions [8].

Quality per cycle distribution, see the Figure 1. To be sure of good sequencer performance one should not see random quality deeps or peaks per cycle. It should gradually decline with cycle because of declining signal-to-noise ratio, as one can see in the Figure 1.

Proportion of duplicate reads should be low, not more than around 10%. Duplicate reads, arising due to PCR (when a library is of low complexity) and optical problems (at the stage of sequencing itself on machine), can bias data towards artificially frequent reads and lead to over-estimating a particular variant contribution in the data. Duplicate's removal is also discussed [33,34].

Proportion if adaptors should be very low, less than 3% at least. Adapter parts might be erroneously sequenced in the beginning of a read, and thus may introduce artificial mutations [35-37]. There are a number of popular tools for adaptor removing from raw sequence reads [34,38-40].

Di-multiplexing, namely separating samples based on their tags, ideally should be even across tags. In pooled *multiplex sequencing*, the size of each pool is a critical issue [38,39]. Thus reasonably even di-multiplexing [20,35,40] ensures less biased data.

There is a good metric for Illumina performance: proportion of purity-filtered data. High purity data is a high signal-to-noise data; see section 3.1 further for more explanations about purity and its filtering. It should be majority of the data, more than 80% in average.

There is also an option to QC check a library even before sequencing. Thus MiSeq QC [41] allows performing a QC run on libraries before deep-sequenced on a larger machine, HiSeq or HiSeqX.

Most sequencers [42,43] generate a QC reports, as part of their processing pipeline. These reports cover a general performance of the corresponding sequencer itself only. They usually do not consider any effects of library preparations and sample extraction.

An exception is FastQC [44]. It is designed to detect problems which originated either in sequencer or in the library preparation step. It is supported by visualising plots and warnings about uncertain results. It is one of the most popular raw sequence reads QC currently. It is a very rapid estimation of various metrics based on stratified sampling of the data.

In contrast, FaQCs [45] monitors errors in the complete bulk of the data, and removes low Q-value reads. Its interesting feature is a k-mer profiling (distribution), which might be of special use for further assembly.

The NGS QC Toolkit [46], besides performing a quality check and providing descriptive statistics, filters low Q data and trims low Q ends of reads. In addition, it allows the conversion between different file formats of NGS data from Illumina and Roche 454 platforms.

Problems and best practices to solve them in QC cleaning

We would recommend filtering in moderation in order not to raise false negative rate. However, any deviation from expected values for the QC metrics mentioned before might be a potential artefact.

Sequence-specific errors create an additional challenge for QC cleaning [47,48]. These are particular combinations of nucleotide, which are prone to signal to noise decline, and therefore are hard to be sequenced, e.g. 'GGT', 'GGC' patterns for Illumina [30,47,48]. The genome regions containing these errors can be covered very unevenly: they can be under-represented in the data, because their quality is usually low, and they are prone to be trimmed out. Moreover, they can be over-represented as well when error prone reads cluster in low complexity repetitive regions.

Aligning Reads to a Reference Genome and/or Assembly Reads

The next step is the matching of the reads to locations at the reference genome, so called mapping. This is done by aligning reads to stretches of the reference genome to which they are most similar in terms of nucleotide sequence. Mapping is the most time and computer memory consuming step [49,50]. It is also crucial: Any artefact in alignment will be subject to further processing and hence propagate errors to the subsequent stages of sequencing.

Because of NGS massive amount of short reads, it is too slow to use the well-known BLAST [51] algorithm, and therefore specific memory and time optimised aligning algorithms, NGS aligners, are developed.

Aligners for NGS vary with respect to their methods, computer resource usage and sensitivity [8,52]. Therefore they may result in different mapping results. It should be mentioned here that the mappability depends on the read length and varies across the genome depending on DNA complexity. Thus sensitivity might be advisable to compute for capture regions in target sequencing, when comparing aligners.

Aligning algorithms also differ in their ability to deal with particular sequencing platforms, protocols, quality of base, and in the handling of structural features of the DNA subject to sequencing (such as repeating motives, gaps, deletions and insertions of nucleotides).

For comparisons and benchmark tests for aligners see [53-57] and the excellent review [50]. The list of aligners is updated online [58].

Box 4: Brief description and classification of NGS aligners and assemblers

NGS alignment algorithms can be divided into two types on the basis of their main methods:

(i) Aligners using hash table indexing method [57]. Typical representatives are MAQ [22], SOAP [59], Novoalign [60], SSAHA [61] etc.;

(ii) Aligners using a Burrows-Wheeler Transform (BWT), such as BWA [62], Bowtie [63], SOAP2 [64] etc.). All aligners usually pre-

process and index both reference and/or reads before actual search of matching read position in the reference genome. A hash table is a type of look up table with more advanced structure of indexing. BWT compresses data in a specific way (modification of a suffix array) before matching. Burrows-Wheeler Transform aligners are faster and use less memory than hash table methods, but are less sensitive [65].

Current assembling algorithms for NGS consist of two main types [66]: (i) Overlap-layout-consensus (OLC) methods; and (ii) Eulerian/de Bruijn Graph (DBG) methods. Both types utilise a graph theory to represent NGS data, but OLC considers reads to be nodes, while DBG takes k-mer for a node. A graph's edges are represented by nodes' overlapping sequences for both classes.

Originally, OLC assemblers were used for longer read Sanger sequencing [67]. They were adopted for shorter reads, e.g. Newbler for 454 and Ion Torrent. However, for larger amount of shorter reads, such as Illumina, SOLiD, DBG methods (Velvet, etc.) became more suitable.

A sequence assembly refers to aligning and merging short fragments from a DNA sequence in order to reconstruct the original sequence. Genome assembling is based on assumption that similar reads belong to the same genomic location [54], which enables to reconstruct genomes after sequencing.

If the genome of a species has not been sequenced before, the assembly of the reads results in the first version of its reference genome. This is called "*de-novo* assembly". Sometimes a *de-novo* assembly is used in combination with alignment in order to reconstruct previously inadequately covered and unreliably sequenced genome regions. It is more expensive computationally compared just to reference-based aligning, but it significantly increases accuracy and completeness of the new assembly/reference [8,52].

Assembling algorithms are very dependent on data types [68], so they closely follow technological developments, and evolve very fast. For an extensive literature on assemblers consult [54,69-73]. The list of aligners is updated online [53].

A comparison/aligning genomes without assembling them is suggested by Patro and Kingsford [74]. There may be an advantage to do so, especially for de novo sequenced genomes. The authors suggested different statistics (based on k-mer distributions within reads) for this comparison. However, possible PCR biases in coverage (and other biases) are not considered at all.

It is not the scope of this paper to review all aligners and assemblers that are currently in use. Here we will concentrate on some of the problems troubling aligners and assemblers in general.

Problems and best practices to solve them: aligners and assemblers

A first stumbling block for somebody wanting to use aligners or assemblers is the sheer number of tools that are available. Thus what aligner or assembly methods to choose becomes not an easy question [68].

Reference errors: One should be aware that an alignment step is obviously dependent on a reference's accuracy. In the case of bad reference, many reference mismatches are not distinguishable from high quality genuine variants. This is also true even in case of an overall well assembled reference (e.g. human) for regions with low mappability. Reference consortium [75] takes care of reliable references.

A bias common to most technologies is that their accuracy decreases with the number of sequential nucleotides identified within a read ("cycle-length") so that errors accumulate at their 3' ends because of error accumulation and molecule degradation [76]. Therefore accuracy of mapping the 3' end of a read suffers. However, one should be careful in clipping the 3' end: The alignment accuracy is affected when read length is short and significant number of bases are clipped [77,78].

Of the more specific shortcomings, we mention platform-dependent issues, complications due to the functional- and structural complexity of the sample DNA and the type of protocol used (reflecting the specific aims of the sequencing endeavour).

Read length and error rates per platform: Depending on the sequencing platform read lengths range from 50-1000 bp [17]. For the main short read platforms the lengths and error rates are as following [17]: Illumina delivers read length 50-300 bp at the error rate 0.1% for end trimmed reads with overall Q30; SOLiD 50-75 bp, 0.1%; Ion Torrent PGM 200-400 bp, 1%; 454 delivers 400-1000 bp reads, 1% error rate.

If reads are short it is more difficult to match them unambiguously to a unique genomic location, because sub-sequences of base typically reoccur many times in a reference genome (this is called "genomic redundancy").

Certain sequencing platforms allow for larger read lengths than others (for example 200 bp by Ion Torrent and 700 bp by Roche's 454 compared to Illumina's reads of 100-250 bp) which makes mapping easier. However, this advantage is outdone by their higher mismatch-error rate; aligners automatically discard reads with too many mismatches on the basis of a pre-set mismatch error rate in a read. Unfortunately, this culling disregards the nature of the mismatch and thus may filter out natural variants.

Sequencing errors is a challenge for aligners [48,79-81]. Obviously, if a read contains more mismatches than allowed by aligner (e.g. 2 per 30 bp seed), than it will not be aligned at all, even if it contains biological signal.

Platform-specific biases: The technology on which a platform is founded may bias it toward particular sequencing mistakes, in turn resulting in platform-specific error profiles.

Some of 'light-based' sequencing platforms, such as SOLiD, Illumina and Complete Genomics, utilise fluorescent dye's labelling to measure signal strength for a corresponding sequencing cycle. These platforms are known to be affected by GC-bias, i.e. a low coverage of either GC-rich or GC-poor (AT-rich) DNA regions [26,82]. It is probably brought about as artefacts of the fragmentation and cloning procedures during library preparation [79,83].

The SOLiD, Illumina and Complete Genomics platforms characteristically suffer from single nucleotide miss-identifications. The SOLiD platform is also known to have problems with sequencing palindromic sequences [84].

Ion Torrent's Personal Genome Machine (PGM) utilises semiconductor sequencing technology that operates on acidity (pH) instead of light. Roche's 454 [85] uses a pyro-sequencing technology. In contrast to the typical single nucleotide miss-identifications of the Illumina and SOLiD, the accuracy of both methods depends on the length of stretches of identical nucleotides, so called homo-polymers. Inaccurate flow-calls result in insertion/deletion (indel) errors, mostly homo-polymer-associated errors, when short homo-polymers are over-

represented, while long are underrepresented [86,87], thus creating an accuracy dependence on a homo-polymer length. Most typical error for the pyrosequencing-type technologies is also indel.

Identifying indels from NGS is known to be very challenging [87], because 'indel by itself interferes with accurate mapping'. To map indels accurately, Pair-End (PE) information is utilised [88]. It works for indels half a size of reads.

Longer deletions are detected by a split-read approach [89], where the information from unmapped reads, likely to contain insertion breakpoints, is utilised. For long insertions a combination with de-novo assembly of poorly covered regions is also required [87]. There is some inconsistency in ranging indel sizes. Here we define short indels are defined as having the size around 1-16 bp, large indels are scaled up to 1 kb, while medium sized are around 16-50 bp [89].

Sequence-specific errors: For pyrosequencing platforms, a 'homopolymer-associated' error leads to discarding repetitive DNA after aligning. There is an evidence of context-dependent indel errors as well. Thus, for Ion Torrent, GC-poor organisms have higher error rate and poorer coverage than GC-balanced [90].

Minoche et al. [80] have shown that for Illumina HiSeq some particular short sub-sequences in plant and virus genomes were accounted for disproportionally high contribution in the overall error rate, up to 24%. This finding was supported by Abnizova et al. [81], where the authors found similar regions with similar motifs associated with high error rate. It was suggested that the origin of this artefact is the motif's vulnerability for Illumina HiSeq common error tendencies: Cross-talk, phasing inaccuracy and G-quenching. Both teams suggested employing a strand -specific quality metric to detect this artefact because of its strand asymmetric distribution.

DNA functionality causes aligning biases: Different parts of a genome are involved in different operations and this is reflected in the nucleotide composition of particular DNA regions. This significantly affects the fragmentation of sample DNA, especially for the whole genome sequencing (G/C splitting bias [91]) as well as the ease by which the fragments can be aligned and mapped.

A study of NGS biases [92], revealed that less complex sequences of introns are less covered with reads (mapped) than more complex sequences of exons. The authors also found that mappability peaks were correlated with biological features, such as intron-exon junction, splice sites, expression level and transcription length.

In line with the above, to confirm an existence of sequencing dependency on DNA functionality, the Auerbach et al. [93] have shown that regions proximal to promoters are prone for sonication breakage, and therefore are the subjects of regional bias. These regions are also responsible for a non-uniform read coverage, producing massive peaks of aligned reads.

One possible solution would be to use UCSC HiSeq Depth tracks [94] where known high sequence depth regions are annotated.

Repetitive DNA causes assembly problem: A particular bothersome feature of the sequential structure of many (if not most) genomes are the presence of large stretches of repetitive DNA (so-called "repeats"): repetitive DNA is consistently overlooked, miss-aligned and miss-assembled by all platforms [28].

More than half of human genome of DNA consists of repetitive elements [95], the fraction of repetitive DNA is even larger for certain plant genomes [96]. Despite of functional importance of repetitive DNA, NGS sequencing often fails to sequence repeats accurately [97,98]. All current technologies are error-biased while dealing with repeats.

But even if a repetitive DNA stretch is sequenced correctly, it might be confused by similar DNA in other genome location, and therefore mis-aligned. In addition, repetitive DNA is often a hot-spot of genuine mutations and structural variations [99].

On a positive side, a lot of the reference genome sequence repeats are already well-known and UCSC tracks [94] can be used to mask these regions. There is evidence [100,101] that single-molecule technologies are helpful in resolving repetitive DNA issues.

Except different repetitive DNA, short indels and segmental duplications are also hard to align [28] because of uncertainty at which location to put an identical DNA stretch.

Box 5: Alu repeat example: For example, a reconstruction of a long mobile elements, such as Alu [102] is still a challenge for a short-read NGS [103]. The reason is that Alu repeats are known to be very abundant in a primate genomes [104]. Though an assemble approaches are successfully utilised for a task of discovery of novel Alu repeats [105,106], the variability of coverage across samples biases the Alu reconstruction towards common insertions.

Assembling is complicated by a repetitive DNA as well. The main assembling assumption (similar reads belong to the same location) is violated by different types of repeats and polymorphic sites. For genomes where the ratio of repeat length to read length is large [54], assembly becomes computationally not tractable. Apparently, if a whole long repetitive stretch were sequenced together with their flanks, it would be easier to locate it back into genome. With longer read technologies there is significant improvement in resolving of repetitive DNA assembly problem [107]. The 10XG linked read sequencing [108], Oxford Nanopore (ONT), PacBio and Illumina TruSeq [28] increased assembly capacities while dealing with repeats [109].

Disadvantage of excessive coverage of repetitive regions (many similar sequences are placed in the same location) can be used creatively. Thus, many software tools for genome mapping and assembly [110,111] uses coverage variability to distinguish unique regions from repetitive ones.

Protocol's diversity: PE and MP usage: Sequencing protocols are very different depending on a researcher's task: e.g. reads sequenced in pairs (pair end, PE and mate-pair, MP) [99,112,113] or singles (SE). PE are utilised by Illumina, and MP by Roche 454 platforms [113]. For Illumina, PE reads help to detect direction and distance between sequenced reads, so reads containing complex DNA can be mapped uniquely [64,114,115].

A special type of PE reads, the long inserts reads (up to 5-10 KB), commonly named as mate-pair libraries [116,117] are useful to link long repeats (including repetitive transposable elements, TEs) and structural variations, and to orient contigs (continues sequences).

Box 6: Single-end, paired-end and mate-pair sequencing: In single-end sequencing (SE), a DNA stretch is sequenced in one direction. In paired-end (PE) sequencing, a DNA stretch is sequenced from both directions. A fluctuation in the expected length between two ends of a PE read after genome alignment can point to a structural variation [118].

Mate-pair is different from PE in library preparation. In PE, the ends of a fragment in a library are sequenced. In contrast to PE, a library for

fragment ends is created for MP first, and then these end fragments are sequenced. In MP sequencing, much longer than for PE, 2-10 kb, fragments are sequenced from both ends. This gives information how far apart nucleotides are linked.

Assembly and mapping problems can be resolved by a longer reads, when it is possible to detect a correct genomic location for a sequenced DNA. Thus, a new synthetic long reads [28] from the Illumina TruSeq are as long as third generation PacBio [100], and has much lower error rate, around 0.03% per base. These long reads are assembled from corresponding Illumina short reads, combining wet lab and computational efforts [119]. Note that the synthetic long reads are essentially single-ended (SE). These reads enable researchers to accurately assemble highly-repetitive TE sequences, such as of a fruit fly genome [119] with a high uniform coverage. However, there are still gaps in assembly reported, together with a low coverage for repetitive GC-low regions (GC-bias).

Unfortunately, as soon as some problems are resolved, side effects of new methods arrive. Thus, the extra-long fragments of mate-pair reads allow discovery of structural variants and de novo assembly. The main problems are: (i) Especially complicated construction of their libraries, and (ii) Frequent mistakes of mapping: 'inward facing' reads instead of 'outward facing', which leads to chimeric read's mapping [41]. Among other problems are: smaller than expected insert sizes [116], AT-rich sequences are underrepresented [120], random spontaneous secondary fragmentation [121].

Assembler's discordance: One more problem is assembler's significant discordance [122]: Different assemblers produce very different amount of assembled data for the same data sets, especially for homologous genome regions.

On a positive side, the 10X Genomics [107] linked-read technology arrived recently, which is capable to assemble thousand Illumina reads into long haplotype phased mega-base blocks [108]. It helps to overcome main problem of short-read aligners and assemblies, namely it allows to locate repetitive regions uniquely. And happily, there is a post-aligning QC option.

Post-Mapping QC

A post-mapping QC is referred to a checking of quality of mapped reads. Mapping is known [123] to be the main source of sequencing artefacts. Prior to the in-depth research analysis it is recommended [124] that one checks the quality of mapped reads because some issues, such as low uneven coverage, homo-polymer biases or experimental artefacts only appear after the alignment.

Problems and best practices to solve them: post-mapping statistics

As one could have noticed from the previous section, there are a lot of challenges for aligners and assemblers. To ensure they worked reasonable, there is a magnitude of QC metrics to track an aligner's performance [124-127]. An amount of post-mapping metrics is large, so we decided to mention several important and/or interesting metrics with brief descriptions, pitfalls and ways to resolve them.

Box 7: List of post-mapping QC tools: The majority of post-mapping QC metrics one can obtain from popular packages such as SAMtools [123], Picard [125], GATK [126], QPLOT [127]. Nice visualization is provided by IGV [128] and GAP5 [129,130]. There are good visualization genome reviews [131,132].

Thus, a *mapping quality score, Q-mapping,* is designed to report a likelihood that a read is aligned correctly [22]. However, a standard output mapping scores of many alignment tools are poorly correlated with actual accuracy of mapping [133]. To solve this problem, a logistic regression method to recalibrate unreliable Q-mapping is suggested by Ruffalo et al. [134]. It is reported to reduce the FP of a downstream variant calling almost 10 times.

Important information on mapped and unmapped read properties is analysed in SAMStat [135]. QC metrics, such as: *Mismatch and indel rates; Insert size distribution; Over-represented k-mers,*-allow to analyse if unmapped read arrived just because of sequencing mistakes, or it contains an important biological signal, e.g splice junction or genomic region escaped from a reference genome.

Proportion of high/low bases and errors: cumulative/survival curves for base calls/errors vs Q-value, see example below:

In the Figure 3, the blue solid line represents cumulative proportion of all bases called with particular Q-value. One can see that there are 90% of bases with Q30+ at the left plot (forward read, R1). The

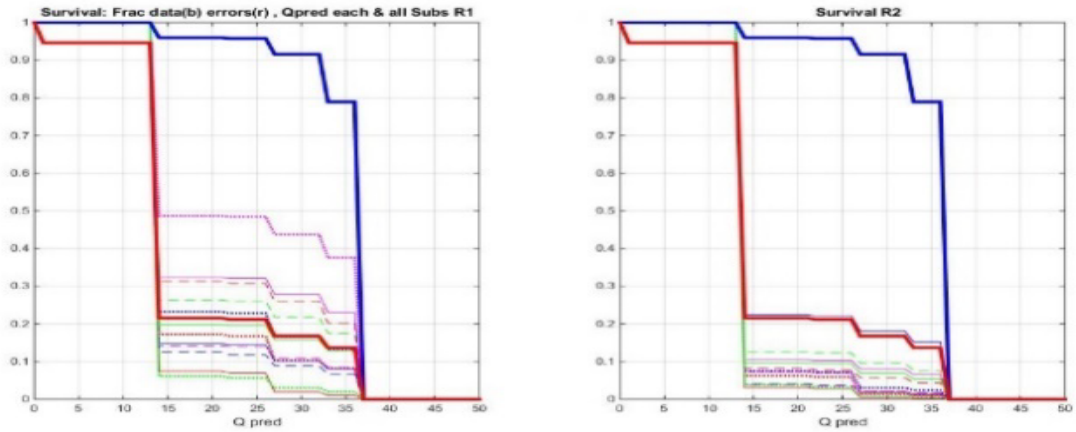

Figure 3: Survival curves for bases (blue) and errors (red). Left- forward read (R1), right-reverse read (R2) in PE. Coloured lines are individual substitution contributions. Illumina HX, Phix genome, sequenced for control at WTSI.

red solid line shows cumulative proportion of all errors and their corresponding Q: there are 18% of error bases with Q30+. Though average error cumulative curves for forward and reverse reads look similar, the contribution of each substitution type is different for each read direction, indicating library preparation artefacts [136].

An interesting metrics is the Genome Mappability Score (GMS) [137]. While mapping score and Q values are assigned to an individual reads, GMS is applied to measure an overall composition of the whole genome. GMS is defined as a weighted probability that any read could be unambiguously mapped to a given location within genome. The measure is used per region, in order to define potentially low-mapped regions based on their complexity. As a result, there are already found large genomic regions, 5-14% of genome (human, mouse, fly, yeast) which are hard to analyse with short reads.

A sequencing coverage depth and evenness are crucial metrics for WGS, enabling accuracy of further analysis, e.g. determination of the confidence of variant calling. Normally, one expects even coverage along a genome, to avoid regional biases. However, coverage is known to be uneven along genome [80], depending on its composition [27], function of a DNA region, and many other features. An excellent paper about theoretical and practical aspects of NGS coverage by Sims et al. [8].

Note that average depth is not very stable metric, it can be easily biased towards overly deep covered regions [10]. These regions (usually repetitive) can be mistakenly pointed for many misaligned repeats. GATK best practices recommend excluding overly deep regions [126]. Median is known to be a more stable measure for coverage because it is less biased towards extreme values.

- Contaminated sequences may introduce artificial mutations [138] when a sample from the same organism are cross-contaminated, for more details read further subsections 5.1.1 and 5.1.2.

- Insert size distribution is a metric of extreme importance for a further data analysis, including variant detection. If it is too different from what was expected by library preparation (for example, too long) it might mean that reads are overlapping, and might introduce FP variant calls [127].

Capture efficiency (the percent of all mapped reads that overlap the targeted regions [139,140] is the most crucial metrics for WES or other target sequencing; it is usually 40-75% for WES. For a more detailed review of the metric Guo et al. research can be approached [10].

Assembly metrics

Box 8: scaffolds and contigs: The result of de novo assembly is a set of DNA strings, called scaffolds. They consist of run of genomic DNA and runs of 'Ns', denoting a gap of estimated length. These 'Ns' are ambiguous bases. The substrings of scaffolds, separated by gaps, are called contigs.

Very useful papers on genome assembly metrics and performance comparison are from Darling et al. [141] and Meader et al. [142]. We want to mention some popular metrics to assess an assembly performance [97,110]:

In the absence of reference genome [143]:

- Number of contigs/scaffolds. The fewer of them the better.
- Contig/scaffolds sizes: max, mean, N50. To calculate a contig N50, one should first re-arrange contigs by ascending order. Then sum up contigs in descending order one by one, starting from the largest contig, until their sum will be equal or more than all contigs total half-sum. The contig N50 of the assembly is the length of the last contig to sum up [144]. The same definition is valid for scaffolds.
- Total size of scaffolds. It should be close to an expected size of a genome assembled.
- Number of Ns (gapped bases) should be as small as possible.

If there is a reference genome, one can assess assembly accuracy and several normalised metrics. Namely normalization takes into account only those parts of assembly that can be aligned to a reference genome using standard local alignment tools. The most popular metrics [145]:

- Sensitivity of assembly is a percent of genome assembled;
- Normalised N50 for contigs;
- Normalised N50 for scaffolds, which is more complicated than for contigs because of N gaps.

Q re-calibration

Even in a raw fastq file before mapping and computing error rates, each base call in a read goes together with its predicted quality, Q-value.

Box 9: A Q-value in a group of a base calls is essentially a log transformation of probability of error of a base call:

$$Q = -10 \times log_{10}(p) \quad (1)$$

Where p is computed as the number of errors divided by the total number of base calls in the group [146].

Brief description: The Q-value/score is the most well accepted measure of base call quality [146,147]. The quality Q-scores compress a variety of types of information about the quality of base calls into a probability-of-error value. Many analysis tools and almost all assemblers and aligners require quality score input to deliver accurate results.

In a raw fastq/bam files these Qs are predicted. The prediction is based on a set of feature values of a base call, and on previous experience with the values of these features. The predicted Q-values are assigned with the help of pre-computed 'canned' look up table, so called calibration table [24,30]. Low Q (Q<20, which corresponds to 0.001 error rate in a group) indicates an ambiguous base call. The high Q (Q>30) usually reflects how successful sequencing on machine was performed and how confident a base call was.

Note that library preparation errors usually have high Q values because they occur before sequencing itself. The errors on a sequencer are usually of low Q, and originate from technological and hardware imperfections. There are well known sources of errors for Illumina sequencers, such as dye label X-talk, phasing inaccuracy [81], molecule degradation with time, G-quenching [148].

Box 10: In Illumina technology, the sequencer analyses one nucleotide of the sequence in each cluster per cycle. At each cycle, A, C, G and T nucleotides, each labelled with a different dye, are added to the flow-cell and the intensity for each dye in each cluster is recorded. Ideally, the strongest of the four intensities recorded at a given cycle for a given cluster should correspond to the nucleotide at that position in the sequence for that cluster. The signal to noise ratio is measured with an index of dominance, called Purity. Two problems accrue in clusters of DNA over successive cycles of the Illumina sequencer, making Purity (and quality) low: phase inaccuracy due to base-incorporation errors and dye-label cross-talk [76]. The phase inaccuracy arises due to base-incorporation errors. A G-quenching is usually lower quality base call

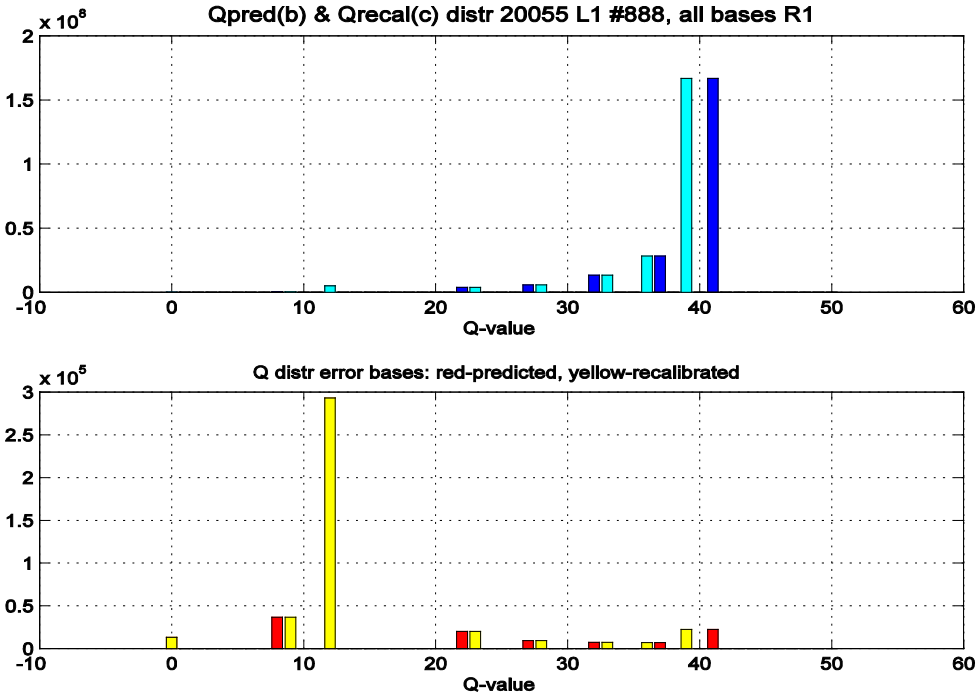

Figure 4: Recalibrated Q-distribution: (top)-all called bases: cyan is for recalibrated Q, blue is for predicted Q, (bottom)-mismatch bases: yellow is for recalibrated Q mismatches, red is for predicted Q of mismatches. Illumina platform, Phix control data.

of a nucleotide preceded [81,148].

Purity is considered high enough, if its value more than 0.6 [30], ensuring strong dominance of intensity of unique base call. Purity filtering is done by computing Purity for each of 25 first base calls in a read. If second maximum Purity of each of base calls is more than 0.6, then the read is defined as a good one. Otherwise the read is filtered out.

It was extremely pronounced for the v3 version HiSeq, and significantly reduced for HiSeqX10 and X5 [136].

The predicted Qs do not always correspond to actual Qs for a particular run/lane/library [148]. In this case (and in case when heterogeneous data are merged) it is recommended to re-calibrate the data using formula (1) [146,149,150]. There is in-house Sanger Institute recalibration and error analysis implemented [30]. The authors attempt not to remove an ambiguous base calls, but to make warnings (low Q) of possible sequencing errors. An example of recalibrated Phix data is shown in the Figure 4: the red and cyan bars are actual Q-scores for errors and all bases, correspondingly.

Reliable Q-value is known to improve SNP call accuracy [151] better than hard filtering. That is why recalibration is recommended as good practice before variant calling.

Variant Identification: Variant Calling and Genotyping

Variant calling from NGS data refers to a computational method for identifying variable sites in genome from the results of NGS experiments [152,153]. Genotype calling determines the genotype for each individual at each site [154].

Though variant identification sounds straightforward and simple: Just compare sequenced samples to a reference genome, in real life it is complicated by various sources of sequencing errors. Thus a good variant caller should compensate or correct for these errors. For a mainstream genotyping and SNP calling guidance we would recommend [155], especially if one wants to analyse human sample. The authors included extra guidance in case of a data not conforming standard assumptions: not having a lot of variance or reliable annotation set.

Variant calling includes small-scale variants [156], such as single nucleotide polymorphisms (SNPs), short insertions and deletions (indels) ranging from 1 to 50 bp in length [157], and large-scale structural variants, Copy Number Variants (CNV) and Structural Variants (SV), which are inversions, translocations, or large indels. Both types of variants relative to a reference are identified by comparison to a reference genome.

Proportion of variation in genomes is significant: e.g. for human genome, SNPs constitute around 0.1%, while SV's impact is estimated as 1.2% [157] and CNV's even as 15% [158].

One of the main uses of NGS is to discover nucleotide level variation between populations of related samples. Thus, variant calling is essential to comparative genomics and genetics of human diseases. An important variant calling application is a clinical testing: Finding disease-associated mutations. Variant calling from NGS can help to detect mutations with a lower frequency than traditionally used Sanger sequencing [159].

For the majority of Whole Genome Sequencing (WGS) and Whole Exome Sequencing (WES) studies detecting of genomic variants is one of the final steps before biological conclusions.

Variant calls are performed in two ways: (1) After aligning reads, or (ii) After assembling. Sometimes these steps are combined. Alignment

of short sequencing reads to a reference genome detects SNPs and small indels in the individuals sequenced, but larger structural variants and repetitive regions in the genome are more difficult to detect. Because structural variation can disrupt genes or regulatory elements, whole-genome sequencing without assembly and detection of structural variation is not complete [123].

For reference -based alignment, the location of each read relative to the reference genome is mapped first. After reads are mapped, a series of QC steps, including duplicate removing, recalibration and indel-realignment, are performed prior to variant calling [160]. The variants are identified by comparison of mapped reads to a reference genome [11,161]. They are further annotated if data about already confirmed variants exist.

For assembly-based alignment, an assembly of raw reads is done first, and only after it this assembly is compared to a reference genome (if it exists). Variant identification after assembling might be useful for individual genes, but becomes less successful when applied to whole genome identification because one cannot use raw reads to verify spurious variants and other genome's contaminations [161].

Box 11: Somatic vs. germline mutations

Variant calling from NGS is successfully applied in genetics of human diseases. There are three common ways how NGS data is used in the area: (a) Identification of causal germline mutations in Mendelian disorders [162,163]; (b) Identification of candidate genes for complex diseases with GWAS [49,164-167]; (c) Identification of somatic and constitutional mutations in cancer [164-168].

It is harder to detect a somatic mutation than a germline mutation [11]. To detect somatic mutations in cancer, Yan et al. and Vissers et al. [169,170] usually compare tumour *vs* normal samples for the same individual.

Box 12: Small list of variant call tools

There are a lot of variant NGS callers, some of them are designed for germline variant calls, others are more suitable for somatic variants. For calling of large-scale structural variant special tools are developed as well. A brief list, far from being complete, of modern variant callers is below:

- Germline variant callers [11] are central for finding rare disease mutations. The most known are: CRISP, GATK [160], SAMtools/bcftools [123], SNver [171-173], VarScan 2 [174].
- Somatic callers are GATK, SAMtools/bcftools, SomaticSniper [175], VarScan 2.
- CNV detection tools are: CNVnator [176], RDXplorer [177], Contra [178], exomeCNV [179]. CNV are usually located within WES or WGS data.
- A SV (inversions, translocations, or large Indels) detection tools are BreakDancer [180], Breakpointer [181], CLEVER [182], SVMerge [159].

The outputs of small-scale variant calls are usually stored in VCF (variant call format) files [183], which are a text comma separated type. An impressive amount of sophisticated statistics is attached for each possible variant position in genome [184]. The large scaled variants are stored in GFF (genetic feature files) format [185]. A machine learning correction is employed by GotCloud [186], where the authors use majority of metrics/features above to train their classifier and correct for biases.

After a variant call it is usually performed an annotation step [64,187,188], followed by visualization [131]. The most common way to annotate is to provide database links to various public variant databases, such as dbSNP. Visual representation can be useful for result's interpretation.

Post-variant Call/Genotyping QC

The list of metrics, which are used as hard filters, is listed below. One can assess a quality of variant calls with these post variant call metrics [186,189-191]. We will mention some of these metrics and their expectation for human NGS data below.

A single nucleotide variants are either transitions, Ti (purine-purine A<->G or pyrimidine-pyrimidine T<->C) or transversions, Tv (pyrine-pyrimidine). The ratio of random changes is expected to be *Ti/Tv*=0.5. However, 1000 G [160] data showed that Ti/Tv for genomic DNA is around 2.1, while for exons it is around 3. Other studies reported that exonic synonymous Ti/Tv is expected as high as 5.6 [192], while other genomic locations are generally in agreement with 1000 G data. Thus Ti/Tv can be used as a measure of variant call quality control.

When analysed 1000G data, Wang et al. [193] found that Ti/Tv negatively correlated with extreme GC-content: The higher Ti/Tv within moderate GC-content (which corresponded to exonic DNA). Hence GC content of the region should be takedn into account during variant calling.

Number of known and novel SNPs per person is estimated to be not more than 200 [194] novel SNPs. *SNP spatial density* can be important parameter. Using data from WGS, the human spatial SNP rate is estimated around 1.1×10^{-8} per site per generation [195]. Viruses have much higher mutation rate, 10^{-3} to 10^{-6} per site per generation [196].

Another metric is heterozygosity ratio: heterozygosity to non-reference homozygosity ratio (het/nonrefhom). It was suggested by Guo et al. [10] to use its value equal 2 as a quality control metric for WGS.

A sequencing coverage depth and evenness are important metrics to determine the confidence of variant calling. The more deep and even is the coverage, the higher is the accuracy of variant calling [8].

To reduce False Positive (FP), many variant callers do a lot of filtering and trimming based on metrics above: using a minimum depth of coverage threshold, base call frequency, masking of homo-polymers and repeats, trimming poor quality bases from a read etc. However, while reducing FP, one can increase false negative (FN) while applying these filters [161,197].

Interestingly, that if one does recalibration of Q-values; it will be comparable and even outperform all hard filtering [186] in respect to variant call accuracy.

To evaluate a performance of a variant caller, one should use a performance metrics (accuracy sensitivity and specificity). These metrics can be derived from contingency tables with FP, FN, TP and TN variant counts [161]. One complication was that there were not a lot of good benchmarking test sets and reliable reference. This complication is partially resolved with selecting a gold standard test set, see further in the text.

There is an ongoing discussion about sensitivity of variant calls for NGS technologies, ranging from very optimistic 100% and recommendations to use NGS for clinical testing [160] (where the variant coverage was in average around 1000x) to much more cautious,

such as 92% sensitivity for SOLiD [26] (with an average coverage around 30x), while 98% for Illumina and Complete Genomics. Thus this coverage difference might explain the difference in accuracy achieved.

A detailed review of post-map QC is done in a research by Guo et al. and Wyllie [190,198]. GATK uses variant QC metrics for their variant calls, using genotyping and known SNP information for a variant QC and annotation. There is a good online resource for a variant detection [199].

Problems and best practices to solve them: variant calling and genotyping

Similarly to aligners and assemblers, there is a large amount of variant call tools now days, so it is hard to choose a right one. Surprisingly, there seems to be lacking a standard evaluation of variant callers [161,200].

There are many complications and challenges for variant detection from NGS data, such as a huge amount of variants. Thus Build 147 provides over 745 million submitted and 250 million reference variants for 7 organisms [201], dbVar [202], and giving large amount of already discovered genetic variations which should be compared with newly called variants during variant annotations.

Large-scale projects, such as 1000 G [203], CHARGE [204] and ExAC [205] significantly contributed into aggregating and harmonising genome and exome sequencing data on genetic human variability, so it became available for the broad scientific community [152] and helps to improve sensitivity of variant detection. One of the conclusions of these projects was that an application of multiple calling algorithms improved the discovery rate and accuracy of genetic variants discovery.

Typical after mapping and after assembling variant call's complications are below:

The most pronounced sources of errors after mapping are repetitive/duplicated genome regions and structural variation [97]. If slightly different regions are mapped to the same locations, they give rise to FP SNP calls. However, these regions are often given low mapping Q, because they have multiple occurrence in genome. In this case they can be filtered out together with true variants, thus given rise to FN. It also should be noted that sequence-specific errors are not always associated with low quality [80,81,206].

There is a special challenge to detect an indel (insertions and deletions) [106] and structural variations [156]. The structural variant call problems are described by Medvedev et al. [99] in details.

Even if reads are mapped correctly, but a region contains small indels or structural variation, it can lead to local misaligning errors, and subsequently to FP and FN calls [207]. For example, reads that aligned at flanking regions of indels, are aligned with mismatches looking as SNP evidence, but are actually alignment artifacts. It is recommended by Van der Auwera et al. [155] to re-align indel locations locally.

Errors because of assembling, such as mis-joined sequences, mis-incorporated adapters etc., produce a false positive variant calls [161,208]. In addition, assemblies often mis-incorporate homo-polymer stretches, specific for some platforms [209], and often collapse multiple alleles into one erroneous variant call. Spurious SNPs are hard to be filtered due to insufficient coverage (1X for assembly), and they are harder to be verified with raw reads. Contaminated genomes are hard to be detected with 'after-assembly' approaches.

Another problem is a reference allele preferential bias, namely that heterozygous alleles might be mapped better for reference variant. In the studies [210,211] this bias was among most important factors which affected allele calling (allele frequency estimation) of targeted sequencing of pooled samples in disease association tests. For the human genome wide study [212] they also found a significant bias toward reference sequence, compared to alternative allele.

In the population genomics studies (such as 1000 G), mapping bias is an important cause of errors in frequency estimation in the HLA (human leucocyte antigen) human genes, reported [213] team. The authors has found that reference allele frequencies were over-estimated in HLA highly polymorphic regions, when analysing 1000 genome data, phase 1.

A genotype calling artefact is still a persistent problem. A visualization of genotype cluster plots for each called SNP is developed by Morris et al. [214] to verify the quality of genotyping.

A discordance/disagreement between variant calling pipelines [215,216] is another serious problem. To overcome this problem, a set of QC metrics to increase reproducibility between SNV-calling pipelines was developed by Wong et al. [159]. These metrics depend only on reference genome used in the alignment without accessing the raw and intermediate data or knowing the SNV calling details.

Another positive development is creating a 'gold standard' SNP training set, based on NA12778 human individual, in order to classify FP and TP in SNP calling by Genome in Bottle (GIAB) consortia [217]. It allowed more systematic comparison of SNP and genotype callers aiming to provide clinical NGS usage guidance [218]. There is growing evidence of improving sensitivity and specificity of variant call by utilizing multiple call consensus approach together with this reliable SNP set [219,220].

Though there are still some drawbacks of machine learning methods to call variants (they are known to be often over-trained, or they might pick up a spurious signal [160,186], the developing of reliable training sets, such as Genome in a bottle data [217], should help to improve reliability of these methods.

Cross-contamination of samples

When variants are called from NGS data, some fraction of low AAF (alternative allele frequency) variants might result because of cross-sample contamination or mix-up of samples [221].

Cross-sample contamination occurs often enough in large-scale NGS studies; its fraction varies from 0.01 to 0.2 depending on a study [221]. It can happen at many steps of a sequencing process: during sample collection, storage, shipping and library preparation. Even if a sample is sequenced without contamination, it can be mixed up computationally at the later stages of merging of multiple runs or di-multiplexing pipeline errors.

There are two major types of contamination: cross-species and within-species. While a cross-species contamination can be filtered out during an alignment, a within species contamination is harder to detect, especially for low coverage studies. Mis-labelling of samples (human error) leads to wrong SNP calls as well.

Contaminated samples often have unusually high levels of heterozygosity [222,223]. It is advised either to exclude contaminated samples from analysis, or model sample contamination during analysis to obtain more accurate SNP and genotype calls.

Cross-individual within species contamination may become a source of FP variant calls, genotype errors and reduced power for association studies. It is very important to distinguish between cross-contamination calls and low level mutation, which is the aim in studies such as cancer somatic mutations (sub-clonal mutations) [152,224].

What is done in the area to solve cross-contamination

Checking for contamination becomes important QC during variant call step, and advised to perform for reliable results [225]. There are several methods to detect cross-sample contamination. However, all of them are usually supported by the additional information about mutations in other samples in a batch [226] or known genotypes [222].

Li and Stoneking [152] detect cross-contamination mutations as minor alleles for contaminated samples which supported by a major allele of some other samples. They also use metrics such as allele balance, depth and recalibrated Q to filter for sequencing artefacts.

For human data the VerifyBamID [222,223] methods is developed, which utilises known SNP information for samples to detect cross-sample contamination. The method suggests supporting contamination estimation by a genotype data.

Recently Flickinger et al. [221] developed a method to estimate and correct for within-species DNA sample contamination during genotyping step. To estimate genotype, they use allele frequencies from the population from which the sample was drawn. Allele frequencies are estimated from a closely related reference population, from array-based genotypes from the same population, or even from the proportion of reads that carry each allele across all sequenced samples.

A brilliant wet-lab based method is developed by Quail et al. [226] for either mixed-up or cross-contaminated samples of any organism. The authors suggested a process where a set of uniquely barcoded DNA fragments are added to samples. From the final sequencing data, one can verify the reads of each sample sequenced and detect contamination or presence of other samples by these barcodes.

To our best knowledge, there is no computational method to estimate cross-contamination for any non-human organism is developed so far. We propose a simple straight-forward method to do so by analysing an alternative allele frequency and depth's distributions directly from an aligned bam data.

NGS Error Models and Simulations

The better we understand and characterise the sources of errors, the better we can cope with them. Thus, it is very useful to derive an error models and to simulate error-prone processes.

It is very important to benchmark existing and newly developed computational methods to process and analyse NGS data. To do so, one can use empirical or simulated data. A pitfall of otherwise valuable real-life empirical data is that genuine process underlying it is often unknown. Therefore it is hard to evaluate accuracy of a method. In contrast, digital simulated data can be generated in controllable way with known parameters. In this way simulated data can complement empirical data for evaluation of a method.

Thus, simulation is used to evaluate performance of bioinformatics tools [227,228], design sequencing projects [229] and computational tools [230]. It is also very useful for evaluating of assemblies [231], gene prediction [232], and genotyping and haplotype reconstruction [154,233].

It is essential to derive good realistic error model for successful simulation of data. An empirically derived, sequence-context based error models are used by Janin et al. [233] to simulate individual sequencing runs and/or technologies. Empirical fragment length and quality score distributions are used. Reads may be drawn from one or more genomes or haplotype sets. In this way one can simulate deep sequencing, meta-genomic, and resequencing. The authors conclude a batch effect: Error profiles can vary noticeably even between different runs of the same NGS technology.

In the interesting study, Orton et al. [234] have developed a computational error model of Illumina's sample processing, including experimental steps. This model predicts which genomic locations are likely to be affected by PCR errors.

However, there is some critique on simulators' redundancy [18].

Discussion

About error correction

Generally speaking, there are two types of error correction: (i) After aligning: An attempt to correct a mismatch between sequenced read and a reference, which looks as an error. (ii) After/during assembling: a consensus (as a majority of base calls) correction of base calls across all reads belonging to the same assembled location, in case at least one of them is different to others.

After alignment step, small amount of mismatches is introduced. Though it is usually a very small proportion of data, they might constitute the most important biological information about sample variability.

However, as one could see before, a significant proportion of mismatches are the result of sequencing and post-processing artefacts and biases. One approach is to compensate for known biases by low Q-score, so they would be warned for further analysis, e.g. error-prone motifs. Another approach is to correct mismatches utilising knowledge about error sources for different platforms' errors [161,235] and computational methods for data processing (aligning, assembling, variant calling).

There are numerous attempts to correct sequencing errors. If sequencing is deep enough (> 10x), they usually correct mismatches by consensus. The authors of Abnizova et al. [81] showed that 80% of errors can be corrected by second best call for Illumina platform, thus reducing FP rate significantly. The authors of Chen et al. [236] proposed to improve sequencing quality and correct for errors by trimming reads for Illumina data. They compared current trimming approaches [46,237] and suggested their alternative one to improve both trimming speed and quality of assembly for trimmed data. Many methods correct for known context biases, such as GGGGT error patterns for Illumina [48].

Box 13: List of error-correction tools: A good list of tools and their short descriptions one can find online [19]. For a platform specific errors there are several error-correction tools such as Musket [238] or Hammer [239] which is performed prior to genome assembly; or PAGIT [240] and Pilon [241] performed after assembly.

A Bayesian approach to correct errors is introduced [242]. It minimizes the error effect on detection low frequency SNPs, applied to pyrosequencing data. There is an attempt to correct for high quality PCR errors in this study.

Another very successful example of correcting PCR errors is

developed by duplex sequencing [243], where the information about forward and reverse complements strands are utilised. In Schulz et al. [244], the correction of indels, especially important for 454 and Ion Torrent data is introduced.

Problems and best practices to solve them in error-correction

An error correction might introduce new type of errors: mis-correction errors [245,246]. And these errors are harder to correct back than technological errors. It might happen that 'averaged' error-corrected repetitive regions are mapped wrongly along genome, creating uneven coverage and false conclusions in further analysis [247,248].

Yang et al. [249] made a sound comparison of NGS platform and a good explanation of current error-correction methods in general and in details. However, the paper's strong points surprisingly lead a reader to a negative conclusion: the paper sounds very convincing, that one should not introduce new mis-correcting errors. From Fujimoto et al. research [245], one can find a conclusion similar to above: error correction methods cannot handle heterozygosity and introduce new mis-corrected errors.

Many methods correct for known context biases, such as GGGGT error patterns for Illumina [80,44]. However, with new Illumina releases (e.g. HXseq or v4), the previous nucleotide dependency is significantly reduced, and new artefacts, such as larger context dependence on a next to mismatch base, arrive [134].

On a positive side, there are successful examples of reduction of error rates and improvement after error correction. Even some time ago there were successful examples of developing strategies to improve reliability of sequencing by machine learning approach [118].

More recently, Wang et al. and Manley et al. [247,248] developed methods which improve the error rate for Illumina sequencing using Phix spike-in external control as an example. Enormous reduction of substitution error rate (93%) for Illumina MiSeq was achieved by Schrimer et al. [249]. They tested different error correction strategies for amplicon sequencing, and found out that quality trimming by Sickle [250] combined with error correction by BayesHammer [239], which then was subjected to read overlapping by Pandaseq [251] gave the best results.

Conclusions

There are definite advantages of NGS compared with Sanger sequencing: the NGS is cheaper and faster than Sanger sequencing [252]. It also has allowed translation of results into clinical practice [253,254].

However, there are still bottle-necks in high-throughput sequencing [255]. One of the main NGS challenges is an overwhelming volume of data generated [252]. It is also challenging to store, analyse [256] and translate this amount of genetic and genomic data into medical and biological context [3,256,257].

Another NGS challenges are dealing with long DNA repeats (because of relatively short reads), and obtaining of uniform genome coverage. Nevertheless, there is the evidence that longer read third generation sequencing is capable to resolve these problems [101,109].

Even more, long synthetic reads facilitate genome phasing [258] and reduce coverage which is needed for phasing [259]. Moreover, the speed of DNA processing already showed how useful long read sequencing can be in medicine and biology. Namely, Oxford Nanopor Technology MinION helped in 2014 Ebola epidemic [260]. One more example is utilising the same device to track the *Salmonela* outbreak in the Stanley Road Hospital, UK [261].

Although second generation sequencing has facilitated SNP and small indels studies at the population level, analysis of larger structural variants is still hard. Third generation mapping and sequencing significantly helped to detect large structural variants [109]. Because many structural variants are surrounded by repeats, the long-range information helps to map repeats and detect variants [137].

The areas mostly beneficial from long read technologies are genome assembly [262,263] and, correspondently, clinical applications [264]. Thus, the authors of [262] assembled several genomes of model organisms (yeast, fruit fly, *Escherichia coli*, Arabidopsis) to very high quality.

Single cell sequencing is hoped to bring more understanding in cell development and disease progressions [144]. We can hope to gain qualitative understanding of life processes in the nearest future.

Acknowledgement

Authors are grateful to Wellcome Trust Sanger Institute and University of Hertfordshire, UK. The work was supported by RFBR (16-54-53064) and ICG SB RAS budget project 0324-2016-0008 (for YLO).

References

1. Robasky K, Lewis NE, Church GM (2014) The role of replicates for error mitigation in next-generation sequencing. Nat Rev Genet 15: 56-62.
2. Sanger F, Nicklen S, Coulson AR (1997) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci 74: 5463-5467.
3. Liu L, Li Y, Li S, Hu N, He Y, et al. (2012) Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012: 251364.
4. Dolled-Filhart MP, Lee M Jr, Ou-Yang CW, Haraksingh RR, Lin JC (2013) Computational and bioinformatics frameworks for next-generation whole exome and genome sequencing. Sci World J 2013: 730210.
5. Schadt EE, Turner S, Kasarskis A (2010) A window into third-generation sequencing. Hum Mol Genet 19: R227-R240.
6. McCourt CM, McArt DG, Mills K, Catherwood MA, Maxwell P, et al. (2013) Validation of next generation sequencing technologies in comparison to current diagnostic gold standards for BRAF, EGFR and KRAS mutational analysis. PLoS One 8: e69604.
7. Clarke LA, Rebelo CS, Gonçalves J, Boavida MG, Jordan P (2001) PCR amplification introduces errors into mononucleotide and dinucleotide repeat sequences. Mol Pathol 54: 351-353.
8. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP (2014) Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 15: 121-132.
9. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C (2014) Ten years of next-generation sequencing technology.Trends Genet 30: 418-426.
10. Guo Y, Ye F, Sheng Q, Clark T, Samuels DC (2014) Three-stage quality control strategies for DNA re-sequencing data. Brief Bioinform 15: 879-889.
11. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, et al. (2014) A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform 15: 256-278.
12. Mutarelli M, Marwah V, Rispoli R, Carrella D, Dharmalingam G, et al. (2014) A community-based resource for automatic exome variant-calling and annotation in Mendelian disorders. BMC Genomics 15: S5.
13. Anders S, Pyl PT, Huber W (2015) HTSe-a Python framework to work with high-throughput sequencing data. Bioinformatics 31: 166-169.
14. Wolfinger MT, Fallmann J, Eggenhofer F, Amman F (2015) ViennaNGS: A toolbox for building efficient next-generation sequencing analysis pipelines.

15. Li JW, Bolser D, Manske M, Giorgi FM, Vyahhi N, et al. (2013) The NGS Wiki Book: a dynamic collaborative online training effort with long-term sustainability. Brief Bioinform 14: 548-555.

16. Li JW, Schmieder R, Ward RM, Delenick J, Olivares EC, et al. (2012) SEQanswers: an open access community for collaboratively decoding genomes. Bioinformatics 28: 1272-1273.

17. Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17: 333-351.

18. Escalona M, Rocha S, Posada D (2016) A comparison of tools for the simulation of genomic next-generation sequencing data. Nat Rev Genet 17: 459-469.

19. Omictools (2016) Error correction software.

20. Hadfield J (2013) Where did It all go Wrong? Quality control for your NGS data.

21. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38: 1767-1771.

22. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18: 1851-1858.

23. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred II error probabilities. Genome Res 8: 186-194.

24. Brockman W, Alvarez P, Young S, Garber M, Giannoukos G, et al. (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res 18: 763-770.

25. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, et al. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 10: R32.

26. Rieber N, Zapatka M, Lasitschka B, Jones D, Northcott P, et al. (2013) Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. PLoS One 8: e66621.

27. Minoche AE, Dohm JC, Himmelbauer H (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol 12: R112.

28. McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, et al. (2014) IlluminaTruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. PLoS One 9: e106689.

29. Ledergerber C, Dessimoz C (2011) Base-calling for next-generation sequencing platforms. Brief Bioinform 12: 489-497.

30. Abnizova I, Skelly T, Naumenko F, Whiteford N, Brown C, et al. (2010) Statistical comparison of methods to estimate the error probability in short-read Illumina sequencing. J Bioinform Comput Biol 8: 579-591.

31. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM (2013) An extensive evaluation of read trimming effects on Illumina NGS data analysis. PLoS One 8: e85024.

32. Kelley DR, Schatz MC, Salzberg SL (2010) Quake: quality-aware detection and correction of sequencing errors. Genome Biol 11: R116.

33. Pireddu L, Leo S, Zanetti G (2011) SEAL: a distributed short read mapping and duplicate removal tool. Bioinformatics 27: 2159-2160.

34. Davis MP, van Dongen S, Abreu-Goodger C, Bartonicek N, Enright AJ (2013) Kraken: a set of tools for quality control and analysis of high-throughput sequence data. Methods 63: 41-49.

35. FASTXtoolkit (2010) FASTQ/A short-reads pre-processing tools.

36. Li JW, Robison K, Martin M, Sjödin A, Usadel B, et al. (2012) The SEQanswers wiki: a wiki database of tools for high-throughput sequencing analysis. Nucleic Acids Res 40: D1313-D1317.

37. Martin M (2011) Cutadapt Removes Adapter Sequences From High-Throughput Sequencing Reads. EMBJ.

38. Marroni F, Pinosio S, Morgante M (2012) The quest for rare variants: pooled multiplexed next generation sequencing in plants. Front Plant Sci 3: 133.

39. Jiang H, Lei R, Ding SW, Zhu S (2014) Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics 15: 182.

40. Mir K, Neuhaus K, Bossert M, Schober S (2013) Short barcodes for next generation sequencing. PLoS One 8: e82933.

41. Illumina (2014) Sequencing Library QC on the MiSeq® System.

42. Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11: 485.

43. http://support.illumina.com/sequencing/sequencing_software/sequencing_analysis_viewer_sav/documentation.html

44. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

45. Lo CC, Chain PS (2014) Rapid evaluation and quality control of next generation sequencing data with FaQCs. BMC Bioinformatics 15: 366.

46. Patel RK, Jain M (2012) NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One 7: e30619.

47. Shin S, Park J (2016) Characterization of sequence-specific errors in various next-generation sequencing systems. Mol Biosyst 12: 914-922.

48. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, et al. (2011) Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res 39: e90.

49. Day-Williams AG, Zeggini E (2011) The effect of next-generation sequencing technology on complex trait research. Eur J Clin Invest 41: 561-567.

50. Fonseca NA, Rung J, Brazma A, Marioni JC (2012) Tools for mapping high-throughput sequencing data. Bioinformatics 28: 3169-3177.

51. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.

52. Mostovoy Y, Levy-Sakin M, Lam J, Lam ET, Hastie AR, et al. (2016) A hybrid approach for de novo human genome sequence assembly and phasing. Nat Method 13: 587-590.

53. Li H, Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11: 473-483.

54. Nagarajan N, Pop M (2013) Sequence assembly demystified. Nat Rev Genet 14: 157-167.

55. Otto C, Stadler PF, Hoffmann S (2014) Lacking alignments? The next-generation sequencing mapper segemehl revisited. Bioinformatics 30: 1837-1843.

56. Pightling AW, Petronella N, Pagotto F (2014) Choice of reference sequence and assembler for alignment of Listeria monocytogenes short-read sequence data greatly influences rates of error in SNP analyses. PLoS One 9: e104579.

57. Shang J, Zhu F, Vongsangnak W, Tang Y, Zhang W, et al. (2014) Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. Biomed Res Int 2014: 309650.

58. http://www.ebi.ac.uk/~nf/hts_mappers/

59. Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: short oligonucleotide alignment program. Bioinformatics 24: 713-714.

60. Novoalign (2014) Novo align NGS quick start tutorial.

61. Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases. Genome Res 11: 1725-1729.

62. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589-595.

63. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.

64. Li R, Yu C, Li Y, Lam TW, Yiu SM, et al. (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25: 1966-1967.

65. https://www.biostars.org/p/11005/

66. Li Z, Chen Y, Mu D, Yuan J, Shi Y, et al. (2012) Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. Brief Funct Genomics 11: 25-37.

67. Seeman T (2011) De novo genome assembly for NGS.

68. Baker M (2012) De novo genome assembly: what every biologist should know. Nat Method 9: 333-337.

69. Chin FYL, Leung HCM, Yiu SM (2014) Sequence assembly using next generation sequencing data-challenges and solutions. Sci China Life Sci 57: 1140-1148.

70. Zhang W, Chen Y, Yang Y, Tang Y, Shang J, et al. (2011) A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. PLoS One 6: e17915.

71. Desai A, Marwah VS, Yadav A, Jha V, Dhaygude K, et al. (2013) Identification of optimum sequencing depth especially for *de novo* genome assembly of small genomes using next generation sequencing data. PLoS One 8: e60204.

72. Wajid B, Serpedin E (2012) Review of general algorithmic features for genome assemblers for next generation sequencers. Genomics Proteomics Bioinformatics 10: 58-73.

73. Schatz MC, Delcher AL, Salzberg SL (2010) Assembly of large genomes using second-generation sequencing. Genome Res 20: 1165-1173.

74. Patro R, Kingsford C (2015) Data-dependent bucketing improves reference-free compression of sequencing reads. Bioinformatics 31: 2770-2777.

75. Wang JM, Zhang K (2015) Microarray analysis of microRNA expression in bone marrow-derived progenitor cells from mice with type 2 diabetes. Genom Data 7: 86-87.

76. Erlich Y, Mitra PP, delaBastide M, McCombie WR, Hannon GJ (2008) Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. Nat Method 5: 679-682.

77. Balint B (2016) Decreased sequencing accuracy at the 3' end of SBS Illumina reads.

78. Sameith K, Roscito JG, Hiller M (2016) Iterative error correction of long sequencing reads maximizes accuracy and improves contig assembly. Brief Bioinform 18: 1-8.

79. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, et al. (2013) Characterizing and measuring bias in sequence data. Genome Biol 14: R51.

80. Minoche AE, Dohm JC, Himmelbauer H (2011) Evaluation of genomic high-throughput sequencing data generated on IlluminaHiSeq and genome analyzer systems. Genome Biol 12: R112.

81. Abnizova I, Leonard S, Skelly T, Brown A, Jackson D, et al. (2012) Analysis of context-dependent errors for illumina sequencing. J Bioinform Comput Biol 10: 1241005.

82. Chen YC, Liu T, Yu CH, Chiang TY, Hwang CC (2013) Effects of GC bias in next-generation-sequencing data on de novo genome assembly. PLoS One 8: e62856.

83. Benjamini Y, Speed TP (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res 40: e72.

84. Huang YF, Chen SC, Chiang YS, Chen TH, Chiu KP (2012) Palindromic sequence impedes sequencing-by-ligation mechanism. BMC Syst Biol 6: S10.

85. Niu B, Fu L, Sun S, Li W (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. BMC Bioinformatics 11: 187.

86. Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW (2013) Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. PLoS Comput Biol 9: e1003031.

87. Li S, Li R, Li H, Lu J, Li Y, et al. (2013) SOAPindel: efficient identification of indels from short paired reads. Genome Res 23: 195-200.

88. Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, et al. (2011) Dindel: accurate indel calls from short-read data. Genome Res 21: 961-973.

89. Kai Ye, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25: 2865-2871.

90. Laehnemann D, Borkhardt A, McHardy AC (2016) Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. Brief Bioinform 17: 154-179.

91. Poptsova MS, Il'icheva IA, Nechipurenko DY, Panchenko LA, Khodikov MV, et al. (2014) Non-random DNA fragmentation in next-generation sequencing. Sci Rep 4: 4532.

92. Schwartz S, Oren R, Ast G (2011) Detection and removal of biases in the analysis of next-generation sequencing reads. PLoS One 6: e16685.

93. Auerbach RK, Euskirchen G, Rozowsky J, Lamarre-Vincent N, Moqtaderi Z, et al. (2009) Mapping accessible chromatin regions using Sono-Seq. Proc Natl Acad Sci U S A 106: 14926-14931.

94. https://genome.ucsc.edu/goldenPath/releaseLog.html

95. Jason de Koning AP, Gu W, Castoe TA, Batzer MA, et al. (2011) Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet 7: e1002384.

96. Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: where genetics meets genomics. Nat Rev Genet 3: 329-341.

97. Alkan C, Sajjadian S, Eichler EE (2011) Limitations of next-generation genome sequence assembly. Nat Methods 8: 61-65.

98. Ye L, Hillier LW, Minx P, Thane N, Locke DP, et al. (2011) A vertebrate case study of the quality of assemblies derived from next-generation sequences. Genome Biol 12: R31.

99. Medvedev P, Stanciu M, Brudno M (2009) Computational methods for discovering structural variation with next-generation sequencing. Nat Methods 6: S13-S20.

100. Sharon D, Tilgner H, Grubert F, Snyder M (2013) A single-molecule long-read survey of the human transcriptome. Nat Biotechnol 31: 1009-1014.

101. Sakai H, Naito K, Ogiso-Tanaka E, Takahashi Y, Iseki K, et al. (2015) The power of single molecule real-time sequencing technology in the de novo assembly of a eukaryotic genome. Sci Rep 5: 16780.

102. Kazazian HH (2004) Mobile elements: drivers of genome evolution. Science 303: 1626-1632.

103. Wildschutte JH, Baron A, Diroff NM, Kidd JM (2015) Discovery and characterization of Alu repeat sequences via precise local read assembly. Nucleic Acids Res 43: 10292-10307.

104. Bennett EA, Keller H, Mills RE, Schmidt S, Moran JV, et al. (2008) Active Aluretrotransposons in the human genome. Genome Res 18: 1875-1883.

105. Chen K, Chen L, Fan X, Wallis J, Ding L, et al. (2014) TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. Genome Res 24: 310-317.

106. Narzisi G, O'Rawe JA, Iossifov I, Fang H, Lee YH, et al. (2014) Accurate de novo and transmitted indel detection in exome-capture data using microassembly. Nat Methods 11: 1033-1036.

107. Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, et al. (2014) Reconstructing complex regions of genomes using long-read sequencing technology. Genome Res 24: 688-696.

108. 10X Genomics (2015) GemCode linked-read platform.

109. Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, et al. (2015) Resolving the complexity of the human genome using single-molecule sequencing. Nature 517: 608-611.

110. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, et al. (2000) A whole-genome assembly of Drosophila. Science 287: 2196-2204.

111. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821-829.

112. GenomeFactory (2013) Paired read confusion.

113. Illumina (2010) Genomic sequencing.

114. Alkan C, Cardone MF, Catacchio CR, Antonacci F, O'Brien SJ, et al. (2011) Genome-wide characterization of centromeric satellites from multiple mammalian genomes. Genome Res 21: 137-145.

115. Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. Genomics 95: 315-327.

116. http://www.illumina.com/documents/products/technotes/technote_nextera_matepair_data_processing.pdf

117. Park N, Shirley L, Gu Y, Keane TM, Swerdlow H, et al. (2013) An improved

approach to mate-paired library preparation for Illumina sequencing. Methods Next Gen Seq 1: 10-20.

118. Kircher M, Stenzel U, Kelso J (2009) Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome Biol 10: R83.

119. Voskoboynik A, Neff NF, Sahoo D, Newman AM, Pushkarev D, et al. (2013) The genome sequence of the colonial chordate, Botryllusschlosseri. Elife 2: e00569.

120. https://gtc.soe.ucsc.edu/content/solid-technology-overview

121. https://lifescience.roche.com/en_gb.html

122. Magoc T, Pabinger S, Canzar S, Liu X, Su Q, et al. (2013) GAGE-B: an evaluation of genome assemblers for bacterial organisms. Bioinformatics 29: 1718-1725.

123. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.

124. GeneStack (2014) Read processing and mapping

125. Picard (2010) Picard tools.

126. GATK (2009) Genome ananysis toolkit.

127. Li B, Zhan X, Wing MK, Anderson P, Kang HM, et al. (2013) QPLOT: a quality assessment tool for next generation sequencing data. Biomed Res Int 2013: 865181.

128. http://www.broadinstitute.org/software/igv/UserGuide

129. http://www.sanger.ac.uk/science/tools/gap5.

130. http://www.sanger.ac.uk/science/tools/categories/sequence-data-processing

131. Nielsen CB, Cantor M, Dubchak I, Gordon D, Wang T (2010) Visualizing genomes: techniques and challenges. Nat Methods 7: S5-S15.

132. Pavlopoulos GA, Malliarakis D, Papanikolaou N, Theodosiou T, Enright AJ, et al. (2015) Visualizing genome and systems biology: technologies, tools, implementation techniques and trends, past, present and future. Gigascience 4: 38.

133. Ruffalo M, LaFramboise T, Koyutürk M (2011) Comparative analysis of algorithms for next-generation sequencing read alignment. Bioinformatics 27: 2790-2796.

134. Ruffalo M, Koyuturk M, Ray S, LaFramboise T (2012) Accurate estimation of short read mapping quality for next-generation genomesequencing. Bioinformatics 28: i349-i355.

135. Lassmann T, Hayashizaki Y, Daub CO (2011) SAMStat: monitoring biases in next generation sequencing data. Bioinformatics 27: 130-131.

136. Abnizova (2017) Changes in context dependencies for new Illumina releases. Manuscript submitted for publication.

137. Lee H, Schatz MC (2012) Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. Bioinformatics 28: 2097-2105.

138. Schmieder R, Edwards R (2011) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One 6: e17288.

139. Garcia-Garcia G, Baux D, Faugere V, Moclyn M, Koenig M, et al. (2016) Assessment of the latest NGS enrichment capture methods in clinical context. Sci Rep 6: 20948.

140. Shearer AE, Hildebrand MS, Ravi H, Joshi S, Guiffre AC, et al. (2012) Pre-capture multiplexing improves efficiency and cost-effectiveness of targeted genomic enrichment. BMC Genomics 13: 618.

141. Darling AE, Tritt A, Eisen JA, Facciotti MT (2011) Mauve assembly metrics. Bioinformatics 27: 2756-2757.

142. Meader S, Hillier LW, Locke D, Ponting CP, Lunter G (2010) Genome assembly quality: Assessment and improvement using the neutral indelmodel. Genome Res 20: 675-684.

143. Simpson JT (2014) Exploring genome characteristics and sequence quality without a reference. Bioinformatics 30: 1228-1235.

144. Stegle O, Teichmann SA, Marioni JC (2015) Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet 16: 133-145.

145. Makinen V, Salmela L, Ylinen J (2012) Normalized N50 assembly metric using gap-restricted co-linear chaining. BMC Bioinformatics 13: 255.

146. Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8: 175-185.

147. Bonfield JK, Staden R (1995) The application of numerical estimates of base calling accuracy to DNA sequencing projects. Nucl Acids Res 23: 1406-1410.

148. https://www.idtdna.com/pages/docs/technical-reports/fluorescence-quenching-by-proximal-g-bases.pdf?sfvrsn=6

149. Whiteford N, Skelly T, Curtis C, Ritchie ME, Lohr A, et al. (2009) Swift: primary data analysis for the IlluminaSolexa sequencing platform. Bioinformatics 25: 2194-2199.

150. Massingham T, Goldman N (2012) All your base: a fast and accurate probabilistic approach to base calling. Genome Biol 13: R13.

151. Li M, Stoneking M (2012) A new approach for detecting low-level mutations in next-generation sequence data. Genome Biol 13: R34.

152. Zhang W, Ng HW, Shu M, Luo H, Su Z, et al. (2015) Comparing genetic variants detected in the 1000 genomes project with SNPs determined by the International HapMap Consortium. J Genet 94: 731-40.

153. Lawrence M (2014) Introduction to variant calling.

154. Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12: 443-451.

155. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, et al. (2013) From fastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics 43: 11.

156. Kojima K, Nariai N, Mimori T, Takahashi M, Yamaguchi-Kabata Y, et al. (2013) A statistical variant calling approach from pedigree information and local haplotyping with phase informative reads. Bioinformatics 29: 2835-2843.

157. Tattini L, D'Aurizio R, Magi A (2015) Detection of genomic structural variants from next-generation sequencing data. Front Bioeng Biotechnol 3: 92.

158. Wong K, Keane TM, Stalker J, Adams DJ (2010) Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. Genome Biol 11: R128.

159. Chin ELH, da Silva C, Hegde M (2013) Assessment of clinical analytical sensitivity and specificity of next-generation sequencing for detection of simple and complex mutations. BMC Genetics 14: 6.

160. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491-498.

161. Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW (2015) Best practices for evaluating single nucleotide variant calling methods for microbial genomics. Front Genet 6: 235.

162. Stitziel NO, Kiezun A, Sunyaev S (2011) Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biol 12: 227.

163. Lettice LA, Hill AE, Devenney PS, Hill RE (2008) Point mutations in a distant sonic hedgehog cis-regulator generate a variable regulatory output responsible for preaxial polydactyly. Hum Mol Genet 17: 978-985.

164. Marian AJ (2012) Molecular genetic studies of complex phenotypes. Transl Res 159: 64-79.

165. Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet 90: 7-24.

166. Lander ES (2011) Initial impact of the sequencing of the human genome. Nature 470: 187-197.

167. Foulkes WD (2008) Inherited susceptibility to common cancers. N Engl J Med 359: 2143-2153.

168. Walther A, Johnstone E, Swanton C, Midgley R, Tomlinson I, et al. (2009) Genetic prognostic and predictive markers in colorectal cancer. Nat Rev Cancer 9: 489-499.

169. Yan XJ, Xu J, Gu ZH, Pan CM, Lu G, et al. (2011) Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocyticleukemia. Nat Genet 43: 309-315.

170. Vissers LE, Fano V, Martinelli D, Campos-Xavier B, Barbuti D, et al. (2011) Whole-exome sequencing detects somatic mutations of IDH1 in metaphysealchondromatosis with D-2-hydroxyglutaric aciduria (MC-HGA). Am J Med Genet A 155A: 2609-2616.

171. Bansal V (2010) A statistical method for the detection of variants from next-generation resequencing of DNA pools. Bioinformatics 26: i318-i324.

172. Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H (2011) SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucleic Acids Res 39: e132.

173. Wang W, Hu W, Hou F, Hu P, Wei Z (2012) SNVerGUI: a desktop tool for variant analysis of next-generation sequencing data. J Med Genet 49: 753-755.

174. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, et al. (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22: 568-576.

175. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, et al. (2012) SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics 28: 311-317.

176. Abyzov A, Urban AE, Snyder M, Gerstein M (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res 21: 974-984.

177. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res 19: 1586-1592.

178. Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, et al. (2012) CONTRA: copy number analysis for targeted resequencing. Bioinformatics 28: 1307-1313.

179. Sathirapongsasuti JF, Lee H, Horst BA, Brunner G, Cochran AJ, et al. (2011) Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics 27: 2648-2654.

180. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, et al. (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 6: 677-681.

181. Sun R, Love MI, Zemojtel T, Emde AK, Chung HR, et al. (2012) Breakpointer: using local mapping artifacts to support sequence breakpoint discovery from single-end reads. Bioinformatics 28: 1024-1025.

182. Marschall T, Costa IG, Canzar S, Bauer M, Klau GW, et al. (2012) CLEVER: clique-enumerating variant finder. Bioinformatics 28: 2875-2882.

183. http://www.1000genomes.org/wiki/Analysis/vcf4.0

184. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, et al. (2011) The variant call format and VCFtools. Bioinformatics 27: 2156-2158.

185. http://www.ensembl.org/info/website/upload/gff.html

186. Jun G, Wing MK, Abecasis GR, Kang HM (2015) An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. Genome Res 25: 918-925.

187. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38: e164.

188. Yang H, Wang K (2015) Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc 10: 1556-1566.

189. Guo Y, Long J, He J, Li CI, Cai Q, et al. (2012) Exome sequencing generates high quality data in non-target regions. BMC Genomics 13: 194.

190. Guo Y, Zhao S, Sheng Q, Ye F, Li J, et al. (2014) Multi-perspective quality control of Illuminaexome sequencing data using QC3. Genomics 103: 323-328.

191. Bainbridge MN, Wang M, Wu Y, Newsham I, Muzny DM, et al. (2011) Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. Genome Biol 12: R68.

192. Gudbjartsson DF, Sulem P, Helgason H, Gylfason A, Gudjonsson SA, et al. (2015) Sequence variants from whole genome sequencing a large group of Icelanders. Sci Data 2: 150011.

193. Wang J, Raskin L, Samuels DC, Shyr Y, Guo Y (2015) Genome measures used for quality control are dependent on gene function and ancestry. Bioinformatics 31: 318-323.

194. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, et al. (2011) Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 12: 745-755.

195. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328: 636-639.

196. Drake JW, Charlesworth B, Charlesworth D, Crow JF (1998) Rates of spontaneous mutation. Genet 148: 1667-1686.

197. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, et al. (2014) Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol 32: 246-251.

198. Wyllie M (2013) Comprehensive analysis of clinical trials data shows unequivocally that Phosphodiesterase Inhibitors (PDEi) improve orgasm. The power of meta-analysis? BJU Int 111: 190-191.

199. Blue Collar Bioinformatics (2013) Framework for evaluating variant detection methods: comparison of aligners and callers.

200. Pabinger S, Trajanoski Z (2013) Genome-scale model management and comparison. Mol Biol 985: 3-16.

201. http://www.ncbi.nlm.nih.gov/SNP/

202. Lin K, Smit S, Bonnema G, Sanchez-Perez G, de Ridder D (2015) Making the difference: integrating structural variation detection tools. Brief Bioinform 16: 852-864.

203. IGSR (2008-2016) 1000 genome project.

204. https://www.dnanexus.com/usecases-charge

205. http://exac.broadinstitute.org/about

206. Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 36: e105.

207. Subramanian S, Di Pierro V, Shah H, Jayaprakash AD, Weisberger I, et al. (2013) MiST: a new approach to variant detection in deep sequencing datasets. Nucleic Acids Res 41: e154.

208. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, et al. (2004) Versatile and open software for comparing large genomes. Genome Biol 5: R12.

209. Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, et al. (2012) High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat Rev Microbiol 10: 599-606.

210. Chen X, Listman JB, Slack FJ, Gelernter J, Zhao H (2012) Biases and errors on allele frequency estimation and disease association tests of next-generation sequencing of pooled samples. Genet Epidemiol 36: 549-560.

211. Guo Y, Cai Q, Li C, Li J, Courtney R, et al. (2013) An evaluation of allele frequency estimation accuracy using pooled sequencing data. Int J Comput Biol Drug Des 6: 279-293.

212. Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, et al. (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics 25: 3207-3212.

213. Brandt DY, Aguiar VR, Bitarello BD, Nunes K, Goudet J, et al. (2015) Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 genomes project phase I data. G3 (Bethesda) 5: 931-941.

214. Morris JA, Randall JC, Maller JB, Barrett JC (2010) Evoker: a visualization tool for genotype intensity data. Bioinformatics 26: 1786-1787.

215. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, et al. (2013) Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med 5: 28.

216. Zhang W, Soika V, Meehan J, Su Z, Ge W, et al. (2015) Quality control metrics improve repeatability and reproducibility of single-nucleotide variants derived from whole-genome sequencing. Pharmacogenomics J 15: 298-309.

217. http://jimb.stanford.edu/giab/

218. Hwang S (2015) Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep.

219. Highnam G, Wang JJ, Kusler D, Zook J, Vijayan V, et al. (2015) An analytical

framework for optimizing variant discovery from personal genomes. Nat Commun 6: 6275.

220. Cornish A, Guda C (2015) A comparison of variant calling pipelines using genome in a bottle as a reference. Biomed Res Int 2015: 456479.

221. Flickinger M, Jun G, Abecasis GR, Boehnke M, Kang HM (2015) Correcting for sample contamination in genotype calling of DNA sequence data. Am J Hum Genet 97: 284-290.

222. Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, et al. (2012) Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet 91: 839-848.

223. Cibulskis K, McKenna A, Fennell T, Banks E, DePristo M, et al. (2011) ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics 27: 2601-2602.

224. Schmitt MW, Loeb LA, Salk JJ, (2016) The influence of subclonal resistance mutations on targeted cancer therapy. Nat Rev Clin Oncol 13: 335-347.

225. http://www.1000genomes.org/

226. Quail MA, Smith M, Jackson D, Leonard S, Skelly T, et al. (2014) SASI-Seq: sample assurance Spike-Ins, and highly differentiating 384 barcoding for Illumina sequencing. BMC Genomics 15: 110.

227. Hu X, Yuan J, Shi Y, Lu J, Liu B, et al. (2012) pIRS: Profile-based Illumina pair-end reads simulator. Bioinformatics 28: 1533-1535.

228. Caboche S, Audebert C, Lemoine Y, Hot D (2014) Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data. BMC Genomics 15: 264.

229. Hoban S, Bertorelle G, Gaggiotti OE (2012) Computer simulations: tools for population and evolutionary genetics. Nat Rev Genet 13: 110-122.

230. Huang W, Li L, Myers JR, Marth GT (2012) ART: a next-generation sequencing read simulator. Bioinformatics 28: 593-594.

231. Knudsen B, Forsberg R, Miyamoto MM (2010) A computer simulator for assessing different challenges and strategies of de novo sequence assembly. Genes (Basel) 1: 263-282.

232. McElroy KE, Luciani F, Thomas T (2012) GemSIM: general, error-model based simulator of next-generation sequencing data. BMC Genomics 13: 74.

233. Janin L, Schulz-Trieglaff O, Cox AJ (2014) BEETL-fastq: a searchable compressed archive for DNA reads. Bioinformatics 30: 2796-2801.

234. Orton RJ, Wright CF, Morelli MJ, King DJ, et al. (2015) Distinguishing low frequency mutations from RT-PCR and sequence errors in viral deep sequencing data. BMC Genomics 16: 229.

235. Edgar RC, Flyvbjerg H (2015) Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics 31: 3476-3482.

236. Chen C, Khaleel SS, Huang H, Wu CH (2014) Software for pre-processing Illumina next-generation sequencing short read sequences. Code Biol Med 9: 8.

237. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114-2120.

238. Liu Y, Schroder J, Schmidt B (2013) Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. Bioinformatics 29: 308-315.

239. Nikolenko SI, Korobeynikov AI, Alekseyev MA (2013) BayesHammer: Bayesian clustering for error correction in single-cell sequencing. BMC Genomics 14: S7.

240. Swain MT, Tsai IJ, Assefa SA, Newbold C, Berriman M, et al. (2012) A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. Nat Protoc 7: 1260-1284.

241. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, et al. (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9: 112963.

242. Zagordi O, Klein R, Daumer M, Beerenwinkel N (2010) Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. Nucleic Acids Res 38: 7400-7409.

243. Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, et al. (2014) Detecting ultralow-frequency mutations by Duplex Sequencing. Nat Protoc 9: 2586-2606.

244. Schulz MH, Weese D, Holtgrewe M, Dimitrova V, Niu S, et al. (2014) Fiona: a parallel and automatic strategy for read error correction. Bioinformatics 30: i356-i363.

245. Fujimoto M, Bodily PM, Okuda N, Clement MJ, Snell Q (2014) Effects of error-correction of heterozygous next-generation sequencing data. BMC Bioinformatics 15: S3.

246. Yang X, Chockalingam SP, Aluru S (2013) A survey of error-correction methods for next-generation sequencing. Bioinform 14: 56-66.

247. Wang XV, Blades N, Ding J, Sultana R, Parmigiani G (2012) Estimation of sequencing error rates in short reads. BMC Bioinformatics 13: 185.

248. Manley LJ, Ma D, Levine SS (2016) Monitoring error rates In Illumina sequencing. J Biomol Tech 27: 125-128.

249. M Schrimer, Ijaz UZ, D'Amore R, Hall N, Sloan WT, et al. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acid Res 43: e37.

250. https://pods.iplantcollaborative.org/wiki/display/DEapps/Sickle-quality-based-trimming

251. Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD (2012) PANDAseq: paired-end assembler for illumina sequences. BMC Bioinformatics 13: 31.

252. Kircher M, Kelso J (2010) High-throughput DNA sequencing--concepts and limitations. Bioessays 32: 524-536.

253. https://www.veritasgenetics.com/documents/VG-launches-999-whole-genome.pdf

254. Schatz MC, Langmead B (2013) The DNA Data Deluge: Fast, efficient genome sequencing machines are spewing out more data than geneticists can analyze. IEEE Spectr 50: 26-33.

255. Pop M, Salzberg SL (2008) Bioinformatics challenges of new sequencing technology. Trends Genet 24: 142-149.

256. Sunyaev SR (2012) Inferring causality and functional significance of human coding DNA variants. Hum Mol Genet 21: R10-R17.

257. Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, et al. (2012) Assuring the quality of next-generation sequencing in clinical laboratory practice. Nat Biotechnol 30: 1033-1036.

258. Snyder MW, Adey A, Kitzman JO, Shendure J (2015) Haplotype-resolved genome sequencing: experimental methods and applications. Nat Rev Genet 16: 344-358.

259. Kuleshov V, Xie D, Chen R, Pushkarev D, Ma Z, et al. (2014) Whole-genome haplotyping using long reads and statistical methods. Nat Biotechnol 32: 261-266.

260. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, et al. (2016) Real-time, portable genome sequencing for Ebola surveillance. Nature 530: 228-232.

261. Quick J, Ashton P, Calus S, Chatt C, Gossain S, et al. (2015) Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. Genome Biol 16: 114.

262. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, et al. (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol 30: 693-700.

263. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, et al. (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 10: 563-569.

264. http://www.nature.com/nrg/collection/clinical-application-next-gen-seq/index.html

Citation: Abnizova I, te Boekhorst R, Orlov Y (2017) Computational Errors and Biases in Short Read Next Generation Sequencing. J Proteomics Bioinform 10: 1-17. doi: 10.4172/jpb.1000420